Title

A simulation-approximation approach to sample size planning for high-dimensional classification studies.

Journal

Biostatistics (Oxford, England), 10(3)

ISSN

1465-4644

Authors

de Valpine, Perry
Bitter, Hans-Marcus
Brown, Michael PS
et al.

Publication Date

2009-07-01

DOI

10.1093/biostatistics/kxp001

Peer reviewed

Biostatistics (2009), **10**, 3, *pp.* 424–435
doi:10.1093/biostatistics/kxp001
Advance Access publication on February 21, 2009

A simulation–approximation approach to sample size planning for high-dimensional classification studies

PERRY DE VALPINE*

Department of Environmental Science, Policy, & Management, University of California, 137 Hilgard Hall No. 3114, Berkeley, CA 94720-3114, USA
pdevalpine@berkeley.edu

HANS-MARCUS BITTER

Roche Palo Alto, 3431 Hillview Avenue, Palo Alto, CA 94304, USA

MICHAEL P. S. BROWN

XDx Expression Diagnostics, 3260 Bayshore Boulevard, Brisbane, CA 94005, USA

JONATHAN HELLER

Predicant Biosciences, 2 N. First Street, San Jose, CA 95113, USA

SUMMARY

Classification studies with high-dimensional measurements and relatively small sample sizes are increasingly common. Prospective analysis of the role of sample sizes in the performance of such studies is important for study design and interpretation of results, but the complexity of typical pattern discovery methods makes this problem challenging. The approach developed here combines Monte Carlo methods and new approximations for linear discriminant analysis, assuming multivariate normal distributions. Monte Carlo methods are used to sample the distribution of which features are selected for a classifier and the mean and variance of features given that they are selected. Given selected features, the linear discriminant problem involves different distributions of training data and generalization data, for which 2 approximations are compared: one based on Taylor series approximation of the generalization error and the other on approximating the discriminant scores as normally distributed. Combining the Monte Carlo and approximation approaches to different aspects of the problem allows efficient estimation of expected generalization error without full simulations of the entire sampling and analysis process. To evaluate the method and investigate realistic study design questions, full simulations are used to ask how validation

*To whom correspondence should be addressed.

error rate depends on the strength and number of informative features, the number of noninformative features, the sample size, and the number of features allowed into the pattern. Both approximation methods perform well for most cases but only the normal discriminant score approximation performs well for cases of very many weakly informative or uninformative dimensions. The simulated cases show that many realistic study designs will typically estimate substantially suboptimal patterns and may have low probability of statistically significant validation results.

Keywords: Biomarker discovery; Experimental design; Generalization error; Genomic; Pattern recognition; Proteomic.

1. Introduction

Recent years have seen an explosion of work on classification problems where the number of measured features per sample is vastly greater than the number of samples. For biological classification problems, such data arise from genomic DNA microarrays and proteomic mass spectrometry assays, from which investigators try to classify disease categories, tumor types, response to drugs, or other categories (Ludwig and Weinstein, 2005). Most of the efforts in method development have appropriately focused on what to do with real data sets (Wang and Shen, 2006; Adam *and others*, 2002). Generally speaking, various methods must select features (sometimes called biomarkers) to be used for classification and estimate a classifier without over-fitting to the many available data dimensions.

Because of the complexity of the algorithms involved, it is not straightforward to answer questions about study design. For example, if there are 10 informative and 5000 noninformative features and the best possible classification error rate is 5%, how many samples are necessary to have an 80% chance of estimating a classifier with less than 10% error rate for independent validation samples? Or, how many samples are necessary so that with probability 95%, the estimated classifier will perform statistically significantly better than a 50% error rate for independent validation samples, that is, conclude the study has at least found something nonrandom? Investigators planning studies have access to sound statistical principles but few specifics to serve as guideposts in evaluating sample sizes relative to hypothesized outcomes. Analysis of study design for high-dimensional classification studies has been identified as an important problem for genomics and proteomics because significant resources are required to execute such studies (Dobbin and Simon, 2007; Allison *and others*, 2006; Pusztai and Hess, 2004; Hwang *and others*, 2002).

Issues of sample size for genomic and proteomic pattern discovery studies are potentially quite important. Over 60 proteomics discovery studies have been published in recent years (Coombes *and others*, 2005; Baker, 2005). Many have sample sizes in the approximately 10–20 range; some notable cases with higher sample sizes (e.g. Adam *and others*, 2002; Petricoin, Ardekani, *and others*, 2002; Petricoin, Ornstein, *and others*, 2002; Zhang *and others*, 2004; Rogers *and others*, 2003) reveal that in broad terms, sample sizes of ~ 50 per group are rare and of ~ 100 per group are very rare. Implicit in some rationales for biomarker discovery studies is the possibility that multiple, individually weak biomarkers could combine to form a collectively strong diagnostic pattern. The observation that discovery studies often find nonspecific markers (Baker, 2005) also suggests that disease-specific patterns may require multiple, individually weak biomarkers. Detecting patterns of multiple weak biomarkers amid many noninformative data dimensions may require substantially greater sample sizes than detecting individually strong biomarkers.

In proteomics, early biomarker discovery and validation studies (Petricoin, Ardekani, *and others*, 2002; Petricoin, Ornstein, *and others*, 2002; Petricoin and Liotta, 2003; Rogers *and others*, 2003; Adam *and others*, 2002; Li *and others*, 2002; Adam *and others*, 2001) led to renewed attention toward potential pitfalls of design and analysis methods. These include low discovery and validation sample sizes, uncertainty about data preprocessing and statistical methods, low sample processing and measurement reproducibility within and between study sites, uncertainty about the biological nature and consistency

of patterns, and lack of independent validation studies (Sorace and Zhan, 2003; Diamandis, 2004a,b; Listgarten and Emili, 2005; Coombes *and others*, 2005; Ebert *and others*, 2006; Wilkins *and others*, 2006). Similar issues have been raised for genomic studies (e.g. Pusztai and Hess, 2004; Ludwig and Weinstein, 2005). Two important studies notable for their independent validation trials highlight the possibility—among many possible reasons for low validation success—that small sample sizes have been fundamentally limiting. Rogers *and others* (2003) saw sensitivity for renal cancer decline from $\sim 100\%$ in discovery to $\sim 40\%$ in validation, and Zhang *and others* (2004) saw specificity decline from $\sim 90\%$ in discovery to $\sim 65\%$ in validation.

For prospective analysis of pattern discovery study designs, purely simulation approaches quickly become cumbersome because there are many scenarios of interest, but purely analytical results are not easy to obtain. We take a middle road between simulations and approximations, with Monte Carlo methods for the feature-selection step and approximations for generalization error rates given each feature set. We use multivariate normal data and linear discriminant classification of features selected by univariate tests. While biologically simplistic, this framework captures the key impacts of both inaccurate feature selection and inaccurate classifier estimation. Related studies that use multivariate normal models include Pepe *and others* (2003), Hu *and others* (2005), Jung (2005), and Dobbin and Simon (2007), among others. Our approach gives order-of-magnitude faster estimation of generalization error compared to direct simulations, which are given for comparison. Both full simulation and simulation–approximation results are useful, but the latter can facilitate more practical exploration of study designs. Our approach also gives insight into which sources of variation are most important and suggests directions for future improvements.

We evaluate the simulation–approximation approach by comparing it to complete simulations that address meaningful study design questions (supplementary material available at *Biostatistics* online, http://www.biostatistics.oxfordjournals.org). We ask how validation error rate depends on the strength and number of informative features (and hence the minimum possible error rate), the number of noninformative features, the patient sample size, and the number of features allowed into the pattern. We find that typical sample sizes may perform poorly when there is a true pattern composed of many individually weak features. This result is not surprising based on general principles, but moving from principles to specific examples as guideposts is important for design of real studies.

We also give 2 approximations of the generalization (or test, or validation) error of a linear discriminant classifier when the training and validation samples do not follow the same distributions. The first is a delta approximation, from Taylor expansions of generalization error around the expected discriminant boundary. The second, and more successful, approximates the discriminant scores as normally distributed. Approximations of linear discriminant analysis with training and generalization samples from the same distributions have been reviewed by McLachlan (1992) and Wyman *and others* (1990). According to Wyman *and others* (1990) and Viollaz *and others* (1995), normal approximations of discriminant scores seem to be more accurate than other approaches, consistent with our results.

A related approach was given by Dobbin and Simon (2007), but ours appears to be more general and accurate (at the expense of being more computational). Theoretical bounds on generalization error from machine learning theory give another path of investigation (Hastie *and others*, 2001). For the related goal of identifying individually significant data dimensions (features), much study design work has built on feature-by-feature false discovery rate ideas (Benjamini and Hochberg, 1995; Storey, 2002; Efron, 2007). Feature-by-feature metrics of study design efficacy include the expected discovery rate (Gadbury *and others* 2004), anticipated average power (Pounds and Cheng, 2005), expected number of false discoveries (Tsai *and others*, 2005), and probability of informative features ranking highly (Pepe *and others*, 2003). Numerous recent studies give methods for feature selection or estimation of generalization error given real data, as opposed to prospective study design (e.g. Mukherjee *and others*, 2003; Fu *and others*, 2005; Wang and Shen, 2006).

2. Problem definition

Consider samples of size n_j for each of J classes ($j \in \{1, \ldots, J\}$), with each sample having M dimensions. By a high-dimensional classification problem, we mean $M \gg n$, where $n = \sum_{j=1}^{J} n_j$ is the total sample size. For the training samples, from which the classifier will be estimated, let $\mathbf{x}_{ij} \in \mathbb{R}^M$ be the data vector for the ith sample of class j. Let $\mathbf{X}_j$ be all the data for class j and $\mathbf{X}$ be all the training data.

Let the number of dimensions of the data distributions that are truly informative (i.e. differ between classes) be M_{I} and those that are truly uninformative be M_{U}, with $M = M_{\mathrm{I}} + M_{\mathrm{U}}$. In the examples below, we will for simplicity use $J = 2$ and group means centered around 0 with all variances equal to 1. Let $\boldsymbol{\Delta}$ be the vector of differences between class means for the informative dimensions, so the means from group 1 are $\left(-0.5\boldsymbol{\Delta}, \mathbf{0}_{M_{\mathrm{U}}}\right)$ and the means from group 2 are $\left(0.5\boldsymbol{\Delta}, \mathbf{0}_{M_{\mathrm{U}}}\right)$, where $\mathbf{0}_{M_{\mathrm{U}}}$ is a length M_{U} vector of zeros. In this notation, a true pattern is defined by $(\boldsymbol{\Delta}, M_{\mathrm{U}})$ and a study design scenario is defined by $(\boldsymbol{\Delta}, M_{\mathrm{U}}, \mathbf{n})$, where $\mathbf{n} = (n_1, n_2)$.

A classifier $\psi(\mathbf{x}_G|\mathbf{X})$ predicts the class, $j \in 1, \ldots, J$, of a new (generalization or validation) sample $\mathbf{x}_G$ based on the training data, $\mathbf{X}$. The generalization sample comes from one of the same distributions (for its unknown class) as the training samples. Define the conditional generalization error for class j as the expected fraction of incorrect classifications for a new sample, $\mathbf{x}_{Gj}$, from class j given a training sample $\mathbf{X}$,

$$\mathrm{CG}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X}) = E_j[I(\psi(\mathbf{x}_{Gj}|\mathbf{X}) \neq j)], \tag{2.1}$$

where the expectation is over $\mathbf{x}_{Gj}$ sampled from true distribution j and the indicator function $I(\mathcal{B})$ is 1 if $\mathcal{B}$ is true and 0 otherwise.

Define the conditional generalization error across all classes as

$$\mathrm{CG}(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X}) = \sum_j P(j)\mathrm{CG}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X}), \tag{2.2}$$

where $P(j)$ is the probability that a new sample is from class j.

The generalization error for a new sample from group j is the conditional generalization error averaged over training samples:

$$G_j(\boldsymbol{\Delta}, M_{\mathrm{U}}, \mathbf{n}) = E_{\mathrm{T}}[\mathrm{CG}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X})], \tag{2.3}$$

where E_{T} denotes expectation over training samples, $\mathbf{X}$, with sample sizes $\mathbf{n}$. Finally, the overall generalization error is

$$G(\boldsymbol{\Delta}, M_{\mathrm{U}}, \mathbf{n}) = \sum_j P(j)G_j(\boldsymbol{\Delta}, M_{\mathrm{U}}, \mathbf{n}). \tag{2.4}$$

Given a generalization sample $\mathbf{X}_G$, with replicate data $\mathbf{x}_{Gj}$ from groups $j = 1, 2$, and a classification procedure $\psi(\mathbf{x}_G|\mathbf{X})$, define the "pattern discovery power" as the expected probability of rejecting the null hypothesis that the predictions $\psi(\mathbf{x}_{Gj}|\mathbf{X})$ are independent of the true class labels, using an appropriate statistical test, with expectations over both the training and generalization samples. This is the probability that the independent validation step of an entire study concludes that the estimated classifier is at least better than random. This paper focuses on calculating generalization error rather than pattern discovery power, but the latter relates to one of the ultimate judgments about a study—whether something nonrandom has been independently validated—and is represented graphically with the simulation results.

3. Simulation–approximation of generalization error

Next, we give a joint simulation and approximation approach to estimate efficiently the generalization error rates CG_j and G_j for multivariate normal data analyzed with linear discriminant analysis. Define a

partition of the space of $\mathbf{X}$ samples into R nonoverlapping regions, $\Omega_1, \ldots, \Omega_R$, that determine which dimensions of $\mathbf{X}$ are selected to estimate the classifier, that is, the feature selection. Define $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_M)$ to be a vector of 0s and 1s, with $\delta_k = 1$ if dimension k will be used for classification and 0 if not. For all $\mathbf{X} \in \Omega_r$, the same dimensions of $\mathbf{X}$ are used by the classifier (so $R \geqslant 2^M$), so it makes sense to write $\boldsymbol{\delta}$ as a function of Ω_r: $\boldsymbol{\delta}_r \equiv \boldsymbol{\delta}(\Omega_r)$.

The generalization error for class j can be factored as

$$G_j(\boldsymbol{\Delta}, M_{\mathrm{U}}, \mathbf{n}) = \sum_{r=1}^{R} P(\mathbf{X} \in \Omega_r) E_{\mathrm{T}}[\mathrm{CG}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X})|\mathbf{X} \in \Omega_r], \tag{3.1}$$

where $P(\cdot)$ is the probability indicated by its argument.

We develop approximations for $E_{\mathrm{T}}[\mathrm{CG}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X})|\mathbf{X} \in \Omega_r]$ based on the first 2 moments of $P(\mathbf{X}|\mathbf{X} \in \Omega_r)$, the probability density of training data sets given that they lead to feature selection $\boldsymbol{\delta}_r$. This is an expected generalization error given that the training and generalization samples do not come from the same distributions. We use Monte Carlo samples to estimate $P(\mathbf{X} \in \Omega_r)$ and the first 2 moments of $P(\mathbf{X}|\mathbf{X} \in \Omega_r)$, which can be generated efficiently. In what follows, $\Omega \in \{\Omega_1, \ldots, \Omega_R\}$.

In a real analysis, feature selection is intertwined with the problem of how many features to include, which is one type of regularization parameter that may be optimized over data-based estimates of generalization error, such as cross-validation. From the study design point of view, the goal is to provide insight into typical study outcomes under various scenarios. Instead of trying to include optimization of the number of features within each approximation, we calculate the approximation across a range of the feature-selection thresholds. This does not include variation or suboptimality in the feature-selection threshold in our estimates of generalization error distributions, but it does offer insight about the sensitivity of generalization error to the feature-selection threshold, which provides context and builds intuition for interpreting results with real data.

3.1 *Monte Carlo approximation of feature selection*

Next, we show how $P(\mathbf{X} \in \Omega)$ and the mean and variance of $P(\mathbf{X}|\mathbf{X} \in \Omega)$ can be estimated with Monte Carlo methods. In the examples here, we assume feature selection is based on feature-by-feature univariate t-tests, which, when the data dimensions really are independent, makes the analysis optimistic because it "knows" this aspect of the "truth." It is common to use feature-by-feature hypothesis tests to estimate false discovery rates as part of analyzing a high-dimensional study, so this simplification allows our results to stand side-by-side with expected false discovery rates and related ideas in considering study designs.

Consider a single data dimension, k, which may or may not be truly informative, for which δ_k will be 1 if the dimension is selected for the pattern and 0 if not. Let x_{ijk} be the kth dimension of sample i from class j. Let the n_1 and n_2 samples from groups $j = 1$ and $j = 2$, respectively, be normally distributed in dimension k: $x_{i1k} \sim \mathcal{N}(-0.5\Delta_k, \sigma^2 = 1)$, $x_{i2k} \sim \mathcal{N}(0.5\Delta_k, \sigma^2 = 1)$. Suppose the decision to include feature k in classification is based on the P-value of a t-test. One calculates $\hat{\mu}_{jk} = n_j^{-1} \sum_{i=1}^{n_j} x_{ijk}$ for $j = 1, 2$; $s_k^2 = \mathrm{df}_s^{-1} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (x_{ijk} - \hat{\mu}_{jk})^2$, where $\mathrm{df}_s = n_1 + n_2 - 2$ are the degrees of freedom of s_k^2; and $t_k = (\hat{\mu}_{2k} - \hat{\mu}_{1k})/(s_k \sqrt{n_1^{-1} + n_2^{-1}})$. The feature is included if $|t_k| > t_{1-P_{\mathrm{c}}/2, \mathrm{df}_s}$, where P_{c} is a threshold significance level for choosing $\delta_k = 1$ and $t_{1-P_{\mathrm{c}}/2, \mathrm{df}_s}$ is the inverse cumulative t-density at $1 - P_{\mathrm{c}}/2$ with df_s degrees of freedom.

It is equivalent to consider the 2 independent random variables

$$z = \frac{(\hat{\mu}_{2k} - \hat{\mu}_{1k})}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}\left(\frac{\Delta_k}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, 1 \right) \tag{3.2}$$

and $e^2 = s_k^2/\sigma^2 \sim \chi^2_{\mathrm{df}_s}$. Then,

$$P(\delta_k = 1) = \int P(z)P(e^2)I\left(\frac{|z|}{\sqrt{e^2/\mathrm{df}_s}} > t_{1-P_c/2,\mathrm{df}_s}\right)\mathrm{d}z\,\mathrm{d}e^2 \tag{3.3}$$

and

$$E[g(z,e^2)|\delta_k = 1] = \frac{\int g(z,e^2)P(z)P(e^2)I\left(\frac{|z|}{\sqrt{e^2/\mathrm{df}_s}} > t_{1-P_c/2,\mathrm{df}_s}\right)\mathrm{d}z\,\mathrm{d}e^2}{P(\delta_k = 1)}. \tag{3.4}$$

Using $g(z,e^2) = \sigma z\sqrt{n_1^{-1} + n_2^{-1}}$ or $g(z,e^2) = \sigma^2 e^2/\mathrm{df}_s$ in (3.4) gives an estimate of the mean difference between groups 1 and 2 or the within-group variance, respectively, given that the t-test is significant.

Working with the densities of z and e^2 allows more efficient numerical methods to estimate (3.3) and (3.4) than if one worked with the densities of x_{ijk} directly. Next, 2 possible Monte Carlo implementations are given, but a variety of numerical methods could be used. For the case of a t-test, (3.3) is simply a cumulative density of a noncentral t-distribution with noncentrality parameter $\Delta_k/(\sigma\sqrt{n_1^{-1} + n_2^{-1}})$ and df_s degrees of freedom. For a Monte Carlo estimate of (3.4), define $\{z^{(l)}, e^{2,(l)}\}$, $l = 1,\ldots,m$, to be a simulated sample from $P\left(z, e^2|I\left(|z|\sqrt{\mathrm{df}_s/e^2} > t_{1-P_c/2,\mathrm{df}_s}\right)\right)$, which can be generated efficiently with a Markov chain Monte Carlo (MCMC) algorithm. Then, a Monte Carlo estimate of (3.4) is

$$\hat{E}[g(z,e^2)|\delta_k = 1] = \frac{1}{m}\sum_{l=1}^{m} g(z^{(l)}, e^{2,(l)}). \tag{3.5}$$

Even a small sample (by MCMC standards) of say $m = 100$ can be reasonable for (3.5).

If one chose to extend the basic idea here for a test for which values of (3.3) are not as easily available as a noncentral t-distribution, then both (3.3) and (3.4) could be estimated by Monte Carlo. For that case, redefine $\{z^{(l)}, e^{2,(l)}\}$, $l = 1,\ldots,m$, to be a Monte Carlo sample of size m from $P(z,e^2)$. Then, the natural estimates of (3.3) and (3.4) are

$$\hat{P}(\delta_k = 1) = \frac{1}{m}\sum_{l=1}^{m} I\left(\frac{|z^{(l)}|}{\sqrt{e^{2,(l)}/\mathrm{df}_s}} > t_{1-P_c/2,\mathrm{df}_s}\right) \tag{3.6}$$

and

$$\hat{E}[g(z,e^2)|\delta_k = 1] = \frac{1}{m\hat{P}(\delta_k = 1)}\sum_{l=1}^{m} g(z^{(l)}, e^{2,(l)})I\left(\frac{|z^{(l)}|}{\sqrt{e^{2,(l)}/\mathrm{df}_s}} > t_{1-P_c/2,\mathrm{df}_s}\right). \tag{3.7}$$

Extensions based on other Monte Carlo numerical integration techniques (such as importance sampling) are straightforward and not our focus here.

3.2 *Approximations for generalization error*

Let $\boldsymbol{\Theta}$ be the parameter vector of the classification function ψ, a linear discriminant function in the examples here. An estimated classifier $\psi(\mathbf{x}_{Gj}|\mathbf{X})$ is defined by estimated parameters $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Theta}}(\mathbf{X})$. For more concise notation, we view generalization error as a function of $\hat{\boldsymbol{\Theta}}$, that is, $\mathrm{CG}_j(\boldsymbol{\Delta}, M_U|\mathbf{X}) = \mathrm{CG}_j(\hat{\boldsymbol{\Theta}})$.
Delta approximation. A delta approximation for the class generalization error given $\mathbf{X} \in \Omega$ is

$$E_T[\mathrm{CG}_j(\hat{\boldsymbol{\Theta}})|\mathbf{X} \in \Omega] \approx \mathrm{CG}_j(E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]) + 0.5\sum_{r,s=1}^{p}\mathrm{Cov}(\hat{\Theta}_r, \hat{\Theta}_s|\mathbf{X} \in \Omega)\partial_{rs}\mathrm{CG}_j(E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]), \tag{3.8}$$

where $\partial_{rs}(\mathrm{CG}(E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]))$ is the second derivative of CG_j with respect to Θ_r and Θ_s evaluated at $E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]$, $\mathrm{Cov}(\hat{\Theta}_r, \hat{\Theta}_s|\mathbf{X} \in \Omega)$ is the covariance between the r and the s dimensions of $\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega$, and p is the number of features selected due to $\mathbf{X} \in \Omega$. The delta approximation is derived by Taylor series expansion of the expectation integral around $E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]$. Note that although the dimensions (or features) are assumed to be independent for feature selection, after they are selected they are approximated as multivariate normal, so the covariances in (3.8) are not necessarily zero.

Normal score approximation. Classifiers typically involve a continuous score function, $d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})$, with prediction of group 1, $\psi(\mathbf{x}_{Gj}|\mathbf{X}) = 1$, if $d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}}) < 0$ (by convention here) and prediction of group 2, $\psi(\mathbf{x}_{Gj}|\mathbf{X}) = 2$, if $d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}}) > 0$. The normal score approximation is to treat $d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})$ as normally distributed with mean $E[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]$ and variance $V[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]$. Then,

$$E_{\mathrm{T}}[\mathrm{CG}_j(\hat{\boldsymbol{\Theta}})|\mathbf{X} \in \Omega] \approx \Phi\left(\frac{u_j E[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]}{\sqrt{V[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]}}\right), \tag{3.9}$$

where $u_1 = +1$, $u_2 = -1$, and $\Phi(\cdot)$ is the standard normal cumulative density function.

Relation to linear discriminant theory. For the case that ψ is a linear discriminant function, we need to calculate $E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]$ and $\mathrm{Cov}(\hat{\Theta}_r, \hat{\Theta}_s|\mathbf{X} \in \Omega)$ for the delta approximation and $E[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]$ and $V[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]$ for the normal score approximation. Define $\mathbf{x}_{Fij}$ to be the selected training features (i.e. given $\mathbf{X} \in \Omega$) of the ith sample from class j. It is convenient to arrange the signs of the data in a consistent manner, so we assume (without loss of generality) that whenever dimension k is included in the classifier, $\hat{\mu}_{2k} > \hat{\mu}_{1k}$ (i.e. if $\hat{\mu}_{2k} < \hat{\mu}_{1k}$, reverse the signs of the data). Then, define $\boldsymbol{\mu}_{Fj}$ and $\boldsymbol{\Sigma}_F$ to be the mean vector and covariance matrix of $\mathbf{x}_{Fij}$, respectively. The difference between means is $\boldsymbol{\Delta}_F = \boldsymbol{\mu}_{F2} - \boldsymbol{\mu}_{F1}$. By symmetry, $\boldsymbol{\mu}_{F1} + \boldsymbol{\mu}_{F2} = \mathbf{0}$. The distributions of the $\mathbf{x}_{Fij}$ will not typically be normal because they are conditioned on a significant difference between normal sample means, but the approximation below uses exact expressions (see supplementary material available at *Biostatistics* online) for $E[\hat{\boldsymbol{\Theta}}|\mathbf{X} \in \Omega]$, $\mathrm{Cov}(\hat{\Theta}_r, \hat{\Theta}_s|\mathbf{X} \in \Omega)$, $E[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]$, and $V[d(\mathbf{x}_{Gj}|\hat{\boldsymbol{\Theta}})]$ under the assumption that the distributions are normal. The expressions use results of Siskind (1972) on the second moments of inverse Wishart distributions, which are related to the sampling distribution of $\hat{\boldsymbol{\Sigma}}_F^{-1}$. This allows full incorporation of multivariate sampling variability in estimating the linear discriminant classifier and uses the principle that second moment–based approximations derived from normal theory are often reasonable. Thus, there are really 2 approximations happening: an approximation of training features (given they have been selected) as multivariate normally distributed and either the delta approximation or normal score approximation of generalization error.

3.3 *Linear discriminant analysis when the training and validation samples follow different distributions*

As above, define a training sample of $\mathbf{x}_{Fi1} \sim N(\boldsymbol{\mu}_{F1}, \boldsymbol{\Sigma}_F)$, $i = 1, \ldots, n_1$, from class 1 and $\mathbf{x}_{Fi2} \sim N(\boldsymbol{\mu}_{F2}, \boldsymbol{\Sigma}_F)$, $i = 1, \ldots, n_2$, from class 2. Define $\boldsymbol{\Delta}_F = \boldsymbol{\mu}_{F2} - \boldsymbol{\mu}_{F1}$, with $\boldsymbol{\Delta}_F > 0$ in every dimension. Define the "true" parameters of ψ in the linear discriminant case as $\boldsymbol{\Theta} = (\mathbf{w}, \mathbf{a})$, where $\mathbf{w} = \boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Delta}_F$ and $\mathbf{a} = 0.5(\boldsymbol{\mu}_{F1} + \boldsymbol{\mu}_{F2}) = \mathbf{0}$. These are estimated by $\hat{\mathbf{w}} = \hat{\boldsymbol{\Sigma}}_F^{-1}\hat{\boldsymbol{\Delta}}_F$, where

$$\hat{\boldsymbol{\Sigma}}_F = \frac{\sum_{i=1}^{n_1}(\mathbf{x}_{Fi1} - \hat{\boldsymbol{\mu}}_{F1})(\mathbf{x}_{Fi1} - \hat{\boldsymbol{\mu}}_{F1})' + \sum_{i=1}^{n_2}(\mathbf{x}_{Fi2} - \hat{\boldsymbol{\mu}}_{F2})(\mathbf{x}_{Fi2} - \hat{\boldsymbol{\mu}}_{F2})'}{n_1 + n_2 - 2} \tag{3.10}$$

is the pooled unbiased estimate of $\boldsymbol{\Sigma}_F$, $\hat{\boldsymbol{\Delta}}_F = \hat{\boldsymbol{\mu}}_{F2} - \hat{\boldsymbol{\mu}}_{F1}$, $\hat{\mathbf{a}} = 0.5(\hat{\boldsymbol{\mu}}_{F1} + \hat{\boldsymbol{\mu}}_{F2})$, and $\hat{\boldsymbol{\mu}}_{Fj} = n_j^{-1}\sum_{i=1}^{n_j}\mathbf{x}_{Fij}$. This is the setup of standard linear discriminant analysis (McLachlan, 1992).

Define a validation sample from class j as $\mathbf{x}_{Gj} \sim N(\boldsymbol{\mu}_{Gj}, \boldsymbol{\Sigma}_G)$. The discriminant score for a value $\mathbf{x}_G$ is

$$d(\mathbf{x}_G|\hat{\boldsymbol{\Theta}}_F) = \hat{\mathbf{w}}'(\mathbf{x}_G - \hat{\mathbf{a}}) - \log\left(\frac{P(1)}{P(2)}\right) \tag{3.11}$$

with prediction of class 1 for $d(\mathbf{x}_G|\hat{\boldsymbol{\Theta}}) < 0$ and class 2 for $d(\mathbf{x}_G|\hat{\boldsymbol{\Theta}}) > 0$. If the training and validation samples came from the same distributions, then $\mathbf{w}$ and $\mathbf{a}$ would give the optimal discriminant function.

We maintain the generality of the prior log-odds ratio, $\log(P(1)/P(2))$, in the derivations. In the simulations below, we assume $P(1) = P(2)$. These values may be very different for a population screening test, where only a very small fraction is expected to have a disease condition, compared to a problem such as disease classification given disease presence. Consideration of $P(1) \neq P(2)$ is standard in balancing sensitivity and specificity of medical tests.

To use the delta approximation (3.8), we need the first 2 moments of $\hat{\mathbf{w}}$ and $\hat{\mathbf{a}}$ and the derivatives of the generalization error with respect to the elements of $\hat{\mathbf{w}}$ and $\hat{\mathbf{a}}$. To use the normal score approximation (3.9), we need the first 2 moments of $d(\mathbf{x}_G|\hat{\boldsymbol{\Theta}}_F)$. These are given exactly in the supplementary material available at *Biostatistics* online for the approximation that the training samples are normally distributed given that the selected dimensions were individually significant.

3.4 *Summation over feature spaces*

It remains to complete the calculation (3.1) efficiently by combining the Monte Carlo estimates of (3.3) and (3.4) and the approximations (3.8) or (3.9). If the space of features that might be selected is relatively simple, then one might directly enumerate cases where $P(\mathbf{X} \in \Omega)$ is appreciably greater than zero; this is not stated mathematically here. More generally, one can use a Monte Carlo sample from the space of selected features to approximate (3.1).

Let $\{\Omega^{(l)}\}, l = 1, \ldots, m$, be a sample from $P(\Omega) \equiv P(\mathbf{X} \in \Omega)$. Corresponding to each partition piece Ω, there is a distribution $P(\mathbf{X}|\mathbf{X} \in \Omega)$. Since this is characterized by $\boldsymbol{\Delta}_F$ and $\boldsymbol{\Sigma}_F$ (estimated by (3.5)), we denote $P(\mathbf{X}|\mathbf{X} \in \Omega^{(l)}) \approx P(\mathbf{X}|\boldsymbol{\Delta}_F^{(l)}, \boldsymbol{\Sigma}_F^{(l)})$. Then, the Monte Carlo approximation of (3.1) is

$$\hat{G}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}, \mathbf{n}) = \frac{1}{m}\sum_{l=1}^{m} E_{\mathrm{T}}[\mathrm{CG}_j(\boldsymbol{\Delta}, M_{\mathrm{U}}|\mathbf{X})|\boldsymbol{\Delta}_F^{(l)}, \boldsymbol{\Sigma}_F^{(l)}]. \tag{3.12}$$

For feature-by-feature selection as discussed above, the relationship $\boldsymbol{\delta}_r = \boldsymbol{\delta}(\Omega_r), r = 1, \ldots, R$, is one-to-one, so we can identify $P(\boldsymbol{\delta}_r) \equiv P(\Omega_r)$. Then, sampling from $P(\Omega)$ in practice amounts to simulating on a feature-by-feature basis whether each feature is selected.

3.5 *Choice of feature-selection thresholds*

The above simulation and approximation steps require a choice for the P-value cutoff, P_{c}, used for feature selection. In practice, one can consider a range of P_{c}-values based on heuristic considerations to encompass the value of P_{c} that minimizes the expected validation error. In the simulation results here (supplementary material available at *Biostatistics* online, summarized below), the following heuristics perform well. The lower bound P_{L} of P_{c} is set to the value at which the probability of zero true discoveries is 30% because excluding most or all informative features will not lead to good patterns. The upper bound P_{U} of P_{c} is the minimum of 2 values. The first is the P_{c} level at which the probability of including all informative features equals 80%, on the rationale that after including most or all informative features, error rates will only get worse as false features are added. The second is the P_{c} such that the expected number of uninformative features is $N/2 - M_{\mathrm{I}}$, that is, the expected total number of features if all truly

informative features are included should not exceed $N/2$. In scenarios where the second bound was lower than the first, higher P_c would lead to worse validation error rates due to many uninformative features.

3.6 *Summary of simulation–approximation method*

In summary, the simulation–approximation procedure uses the following steps:

1. Choose $\boldsymbol{\Delta}$, M_U, and $\mathbf{n}$ to define a study scenario.
2. Choose a useful range of feature-selection thresholds, P_c, which influence how many features are chosen in the feature-selection stage.
3. For each (unique) dimension of $\boldsymbol{\Delta}$ and each P_c, use the noncentral t-distribution and/or Monte Carlo methods to estimate
    a) the probability that the feature will be selected,
    b) the expected within-group variances and difference between group means given that the feature is selected.
4. For the Monte Carlo approximation (3.12), generate a sample of training feature combinations, $\{\boldsymbol{\Omega}^{(l)}\}$, for which the generalization error will be approximated.
5. For each training feature combination, use the variances and mean differences given that the features are selected to approximate the generalization error using either (3.8) or (3.9) with the calculations in the supplementary material available at *Biostatistics* online.
6. Sum the terms in (3.12).

4. SIMULATION STUDY

Results of simulations of 7 realistic study designs are detailed in the supplementary material available at *Biostatistics* online. The first 6 scenarios consider optimal (i.e. Bayes) error rates of 0.05, 0.10, and 0.20 with either 3 (few strong) or 12 (many weak) truly informative dimensions, while the seventh considers optimal error of 0.05 from 46 (very many, very weak) dimensions. All scenarios use equal discovery sample sizes for control and disease groups, $n_1 = n_2$, with the same mean difference for all informative dimensions and 10 patients per group for validation power. The simulation–approximation is accurate with the normal score approximation in all scenarios and with the delta approximation in all scenarios except for very many, very weak true features. Both methods are most accurate when most of the variation in generalization error is due to variation in which features are selected rather than in discriminant parameters given the feature space. Much larger numbers of truly informative dimensions would render the approximations inaccurate, and, moreover, suggest methods beyond basic linear discriminant analysis (LDA), such as shrinkage methods to constrain high variances in estimated patterns.

Several realistic scenarios have limited statistical power for validation and lead to substantially suboptimal patterns. With 12 informative and 2000 uninformative features and optimal error rate of 20%, sample sizes of 20, 50, and 100 give median validation error rates around 48%, 40%, and 30–35%, respectively, with only sample sizes of 100 giving better than 50% power for validation. If the features give optimal error rate of 10%, then 50 patients per group give high validation power but with median error rates of roughly 18–22% for 1000–5000 uninformative dimensions. With an optimal error rate of 5%, 20 samples would give roughly 50–80% validation power at 5% significance for 1000–5000 uninformative dimensions.

For a given optimal error rate, it is much harder to find patterns from many weak than from few strong informative features. Given optimal error rate of 20%, 50 patients per group for 3 strong features give better results than 100 patients per group with 12 weak features. For optimal error rate of 10% or 5%, 20

patients per group for 3 strong features give roughly comparable performance to 50 patients per group for 12 weak features. In summary, by far the strongest factors in pattern discovery power are sample size and individual feature strength. Some of these results are sobering in light of sample sizes in typical studies. It is plausible that some real studies to discover diagnostic patterns from high-dimensional assays could have low power for independent validation and find patterns far from the best true pattern.

5. Discussion

Prospective analysis of study design for high-dimensional pattern discovery is important to plan studies with reasonable expectations of success based on scientific guesswork about the types of real patterns that might exist. The complexity of feature selection and pattern analysis methods raises many challenges for prospective study design. Here, we have explored a middle road between simulation and approximation, with simulations to handle variability in the selected features and an approximation of linear discriminant analysis given that the selected features appear to be informative in training data.

One of the most complicated ways in which the scenarios here may be optimistic is their lack of multivariate patterns and pattern recognition methods. Multivariate patterns could include correlated features that appear to be individually weak but are collectively strong or even harder possibilities such as the classic "XOR" (checkerboard) problem, where each marginal distribution has no information and only more complicated models than LDA can represent the pattern. In such problems, the hazard of over-fitting is greater than for the simulations here and would likely produce less favorable results. Other directions for further exploration of the relationships between sample size, numbers of informative and noninformative features, true optimal error rate, and discovery and generalization error rates include the following: generation of data from distributions that are unknown to the learning method (i.e. non-normal), further development of the relationship between false discovery rates and pattern discovery power, and further theoretical development of accurate approximations and/or efficient simulations.

Acknowledgments

This work was initiated while all authors were employed at Predicant Biosciences. We thank our colleagues at Predicant for insightful discussions and support. *Conflict of Interest:* None declared.

References

Adam, B. L., Qu, Y. S., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z. D. *and others* (2002). Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* **62**, 3609–3614.

Adam, B. L., Vlahou, A., Semmes, O. J. and Wright, G. L. (2001). Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics* **1**, 1264–1270.

Allison, D. B., Cui, X. Q., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

Baker, M. (2005). In biomarkers we trust? *Nature Biotechnology* **23**, 297–304.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B—Methodological* **57**, 289–300.

Coombes, K. R., Morris, J. R. S., Hu, J. H., Edmonson, S. R. and Baggerly, K. A. (2005). Serum proteomics profiling—a young technology begins to mature. *Nature Biotechnology* **23**, 291–292.

Diamandis, E. P. (2004a). Mass spectrometry as a diagnostic and a cancer biomarker discovery tool—opportunities and potential limitations. *Molecular & Cellular Proteomics* **3**, 367–378.

DIAMANDIS, E. P. (2004b). Proteomic patterns to identify ovarian cancer: 3 years on. *Expert Review of Molecular Diagnostics* **4**, 575–577.

DOBBIN, K. K. AND SIMON, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8**, 101–117.

EBERT, M. P. A., KORC, M., MALFERTHEINER, P. AND ROCKEN, C. (2006). Advances, challenges, and limitations in serum-proteome-based cancer diagnosis. *Journal of Proteome Research* **5**, 19–25.

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102**, 93–103.

FU, W. J., DOUGHERTY, E. R., MALLICK, B. AND CARROLL, R. J. (2005). How many samples are needed to build a classifier: a general sequential approach. *Bioinformatics (Oxford)* **21**, 63–70.

GADBURY, G. L., PAGE, G. P., EDWARDS, J., KAYO, T., PROLLA, T. A., WEINDRUCH, R., PERMANA, P. A., MOUNTZ, J. D. AND ALLISON, D. B. (2004). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research* **13**, 325–338.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

HU, J., ZOU, F. AND WRIGHT, F. A. (2005). Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics (Oxford)* **21**, 3264–3272.

HWANG, D. H., SCHMITT, W. A., STEPHANOPOULOS, G. AND STEPHANOPOULOS, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics* **18**, 1184–1193.

JUNG, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics (Oxford)* **21**, 3097–3104.

LI, J. N., ZHANG, Z., ROSENZWEIG, J., WANG, Y. Y. AND CHAN, D. W. (2002). Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry* **48**, 1296–1304.

LISTGARTEN, J. AND EMILI, A. (2005). Practical proteomic biomarker discovery: taking a step back to leap forward. *Drug Discovery Today* **10**, 1697–1702.

LUDWIG, J. A. AND WEINSTEIN, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer* **5**, 845–856.

MCLACHLAN, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.

MUKHERJEE, S., TAMAYO, P., ROGERS, S., RIFKIN, R., ENGLE, A., CAMPBELL, C., GOLUB, T. R. AND MESIROV, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* **10**, 119–142.

PEPE, M. S., LONGTON, G., ANDERSON, G. L. AND SCHUMMER, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133–142.

PETRICOIN, E. F., ARDEKANI, A. M., HITT, B. A., LEVINE, P. J., FUSARO, V. A., STEINBERG, S. M., MILLS, G. B., SIMONE, C., FISHMAN, D. A., KOHN, E. C. *and others* (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577.

PETRICOIN, E. F. AND LIOTTA, L. A. (2003). Mass spectrometry-based diagnostics: the upcoming revolution in disease detection. *Clinical Chemistry* **49**, 533–534.

PETRICOIN, E. F., ORNSTEIN, D. K., PAWELETZ, C. P., ARDEKANI, A., HACKETT, P. S., HITT, B. A., VELASSCO, A., TRUCCO, C., WIEGAND, L., WOOD, K. *and others* (2002). Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute* **94**, 1576–1578.

POUNDS, S. AND CHENG, C. (2005). Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263–4271.

PUSZTAI, L. AND HESS, K. R. (2004). Clinical trial design for microarray predictive marker discovery and assessment. *Annals of Oncology* **15**, 1731–1737.

ROGERS, M. A., CLARKE, P., NOBLE, J., MUNRO, N. P., PAUL, A., SELBY, P. J. AND BANKS, R. E. (2003). Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. *Cancer Research* **63**, 6971–6983.

SISKIND, V. (1972). Second moments of inverse Wishart-matrix elements. *Biometrika* **59**, 690–691.

SORACE, J. M. AND ZHAN, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **4**, 24.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B—Statistical Methodology* **64**, 479–498.

TSAI, C.-A., WANG, S.-J., CHEN, D.-T. AND CHEN, J. J. (2005). Sample size for gene expression microarray experiments. *Bioinformatics (Oxford)* **21**, 1502–1508.

VIOLLAZ, A. J., SFER, A. M. AND SALVATIERRA, S. M. (1995). An approximation of the unconditional error rates of the sample linear discriminant function. *Communications in Statistics—Theory and Methods* **24**, 1941–1969.

WANG, J. H. AND SHEN, X. T. (2006). Estimation of generalization error: random and fixed inputs. *Statistica Sinica* **16**, 569–588.

WILKINS, M. R., APPEL, R. D., VAN EYK, J. E., CHUNG, M. C. M., GORG, A., HECKER, M., HUBER, L. A., LANGEN, H., LINK, A. J., PAIK, Y. K. *and others* (2006). Guidelines for the next 10 years of proteomics. *Proteomics* **6**, 4–8.

WYMAN, F. J., YOUNG, D. M. AND TURNER, D. W. (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition* **23**, 775–783.

ZHANG, Z., BAST, R. C., YU, Y. H., LI, J. N., SOKOLL, L. J., RAI, A. J., ROSENZWEIG, J. M., CAMERON, B., WANG, Y. Y., MENG, X. Y. *and others* (2004). Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Research* **64**, 5882–5890.

[*Received May 29, 2007; revised August 19, 2008; second revision November 24, 2008; accepted for publication January 20, 2009*]